

# Cluster networks and Bruhat–Tits buildings

S.V.Kozyrev\*

April 29, 2014

## Abstract

Clustering procedure for the case where instead of a fixed metric one applies a family of metrics is considered. In this case instead of a classification tree one obtains a classification network (a directed acyclic graph with non directed cycles).

Relation to Bruhat–Tits buildings is discussed. Dimension of a general cluster system is considered.

Keywords: clustering, cluster networks, Bruhat–Tits buildings.

## 1 Introduction

This paper discusses the relation of construction of multidimensional cluster systems described in [1], [2] and geometry of Bruhat–Tits buildings. Application to data analysis is also discussed.

Clustering procedure gives a construction of a tree of clusters with a hierarchy (partial order) starting from a metric on a set of points. This procedure is an important method in data analysis, in particular in applications to bioinformatics (construction of taxonomy, or a tree of life).

In the present paper we discuss the following approach to clustering: assume we have instead of one metric a family of metrics depending on a set of parameters (this is a typical situation in applications). We will get a family of clusterings. The question is: can we describe this family by a single mathematical object?

We will use an analogy from  $p$ -adic geometry. In  $p$ -adic spaces we have natural hierarchies (partially ordered trees) of balls. A hierarchy of this kind can be considered as a clustering with respect to metric in  $p$ -adic space. In multidimensional  $p$ -adic geometry we have a generalization of hierarchies, described by the affine Bruhat–Tits buildings. These buildings are related to families of balls with respect to several ultrametrics.

In this paper we discuss a relation between Bruhat–Tits buildings and geometry of  $p$ -adic cluster systems with respect to a family of metrics. We also discuss a generalization of the corresponding structure of simplicial complex to

---

\*Steklov Mathematical Institute, Russian Academy of Sciences

general systems of clusters. In particular, we discuss a notion of dimension for general cluster systems.

For other application of  $p$ -adic numbers to data analysis see [3], [4].

The structure of the present paper is as follows.

In Section 2 we recall the definition of clustering procedure and discuss some examples of cluster networks for a family of metrics.

In Section 3 we recall the definition of the affine Bruhat–Tits buildings and discuss the relation of this building and networks of clusters in  $p$ -adic spaces.

In Section 4 we discuss a structure of simplicial complex for general cluster networks and a definition of dimension form general cluster systems.

## 2 Examples of cluster networks

Let us recall the standard definition of clustering. We will use for simplicity the nearest neighbor clustering (we could also consider more general clustering algorithms). For general discussion of clustering see [5].

Let  $(M, \rho)$  be an arbitrary metric space.

A sequence of points  $a = x_0, x_1, \dots, x_{n-1}, x_n = b$  in  $(M, \rho)$  is called an  $\varepsilon$ -chain connecting two points  $a$  and  $b$  if  $\rho(x_k, x_{k+1}) \leq \varepsilon$  for all  $0 \leq k < n$ .

If there exists an  $\varepsilon$ -chain connecting  $a$  and  $b$  then  $a$  and  $b$  are  $\varepsilon$ -connected.

The chain distance between  $a$  and  $b$  is defined as  $d(a, b) = \inf(\varepsilon: a, b \text{ } \varepsilon\text{-connected})$ .

This distance has all properties of ultrametric excluding non-degeneracy (i.e. non coinciding points can have zero chain distance). Therefore chain distance defines an ultrametric on the set of equivalence classes of points in  $M$ , where  $a, b$  are in the same equivalence class when  $d(a, b) = 0$ .

In particular if the initial metric  $\rho$  is an ultrametric, then the corresponding chain distance  $d$  coincides with  $\rho$ .

A cluster  $C(i, R)$  in a metric space  $(M, \rho)$  is a ball with the center  $i$  and radius  $R$  with respect to the chain distance, i.e. the set  $\{j \in M: d(i, j) \leq R\}$ .

A clustering of a metric space  $M$  is a cluster set, satisfying:

- i) every element in  $M$  belongs to some cluster;
- ii) for any pair  $a, b$  of elements in  $M$  there exists a minimal cluster  $\sup(a, b)$  containing both elements;
- iii) for arbitrary embedded clusters  $A \subset B$  every increasing sequence of embedded clusters  $\{A_i\}$ ,  $A \subset \dots \subset A_i \subset A_{i+1} \subset \dots \subset B$  is finite.

Clustering procedure generates a partially ordered tree of clusters (dendrogram) in the following way:

- i) vertices are clusters;
- ii) partial order is given by inclusion of clusters;
- iii) edge connects two clusters nested without intermediaries.

Multidimensional generalization of clustering [1], [2] is introduced as a generalization of clustering for the case of several metrics. Assume we have a family

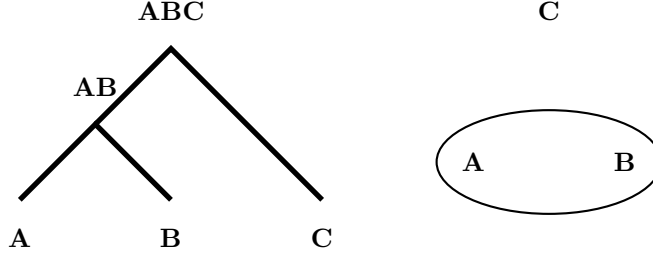


Fig. 1: Cluster tree  $\mathcal{A}_1$

of metrics (which is a standard situation for applications in data analysis), say this family is parameterized by a set of real parameters. In this case instead of one cluster tree we will obtain a family of cluster trees. Some of clusters for the different trees can coincide as sets. Identifying these clusters for the different trees we will obtain a network of clusters. We will discuss the relation of cluster networks of this type and Bruhat–Tits buildings.

Before the discussion of general definition let us consider examples of multi-dimensional clustering.

**Example 1.** The case of a set of three points  $A, B, C$  in a two-dimensional real plane  $\mathbb{R}^2$  with the standard Euclidean metric. Parameters defining the metric are coordinates of the points in the plane.

Cluster tree  $\mathcal{A}_1$  (see Fig. 1). The cluster set contains  $A, B, C, AB, ABC$  (vertices of the cluster tree), edges join the vertices in accordance with the growth of the clusters, i.e. the cluster tree contains the edges

$$(A, AB), (B, AB), (AB, ABC), (C, ABC).$$

Here we denote by  $ABC$  the cluster containing  $A, B$  and  $C$ .

Cluster tree  $\mathcal{B}_1$  (see Fig. 2). Let us consider a deformation of the metric (motion of the points in the plane  $\mathbb{R}^2$ ) with the replacing of the above cluster set by the set of vertices  $A, B, C, BC, ABC$  with the corresponding edges

$$(B, BC), (C, BC), (BC, ABC), (A, ABC).$$

Cluster network  $\mathcal{C}_1$  (see Fig. 3). This network is a union of the trees of clusters  $\mathcal{A}_1$  and  $\mathcal{B}_1$  (where we identify the clusters which coincide as sets).

The vertex set of the network  $\mathcal{C}_1$  contains the clusters

$$A, B, C, AB, BC, ABC$$

the set of edges of  $\mathcal{C}_1$  contains

$$(A, AB), (B, AB), (AB, ABC), (C, ABC), \\ (B, BC), (C, BC), (BC, ABC), (A, ABC).$$

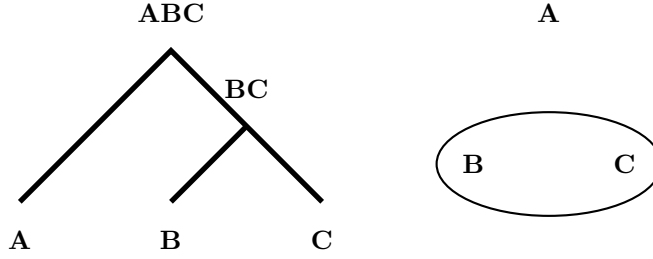


Fig. 2: Cluster tree  $\mathcal{B}_1$

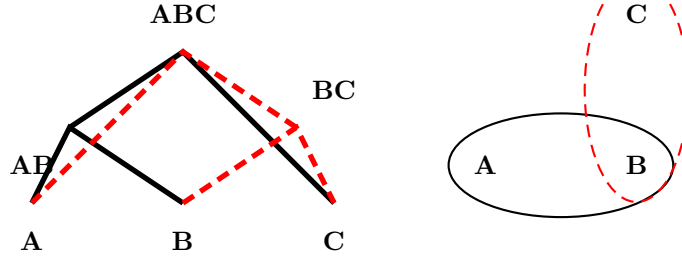


Fig. 3: Cluster network  $\mathcal{C}_1$

The partial order of vertices is given by inclusion of clusters.

Cycles in this graph describe the different histories of the growth of clusters (growth with respect to the different metrics).

**Example 2.** Let us consider a set of four points  $A, B, C, D$  located in the plane  $\mathbb{R}^2$  at the vertices of some quadrangle. We will have the following trees of clusters.

Cluster tree  $\mathcal{A}_2$  (see Fig. 4). Clustering with respect to the plane metric gives the clusters

$$A, B, C, D, AB, CD, ABCD.$$

The set of edges has the form

$$(A, AB), (B, AB), (C, CD), (D, CD), (AB, ABCD), (CD, ABCD).$$

Cluster tree  $\mathcal{B}_2$  (see Fig. 5). Deformation of the mentioned quadrangle gives the cluster set

$$A, B, C, D, AC, BD, ABCD$$

with the edges

$$(A, AC), (C, AC), (B, BD), (D, BD), (AC, ABCD), (BD, ABCD).$$

Cluster network  $\mathcal{C}_2$  (see Fig. 6) is the union of the trees  $\mathcal{A}_2$  and  $\mathcal{B}_2$  of clusters. This network contains the unions of the vertex sets and the edges sets in  $\mathcal{A}_2$  and  $\mathcal{B}_2$ .

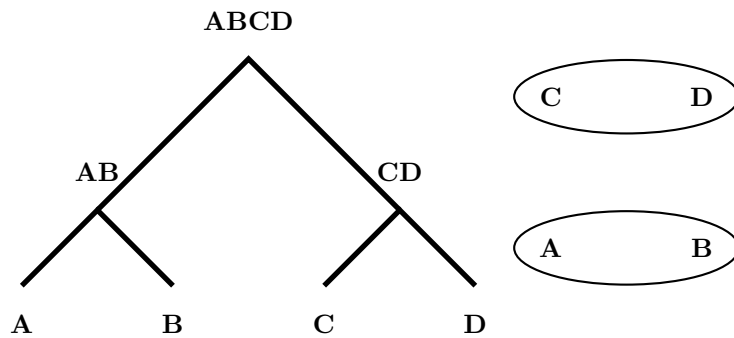


Fig. 4: Cluster tree  $\mathcal{A}_2$

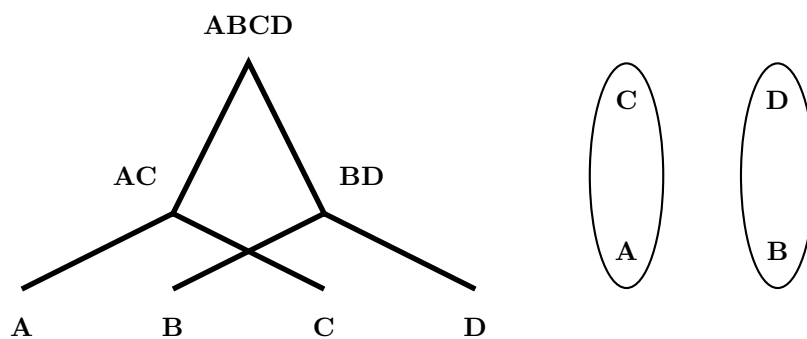


Fig. 5: Cluster tree  $\mathcal{B}_2$

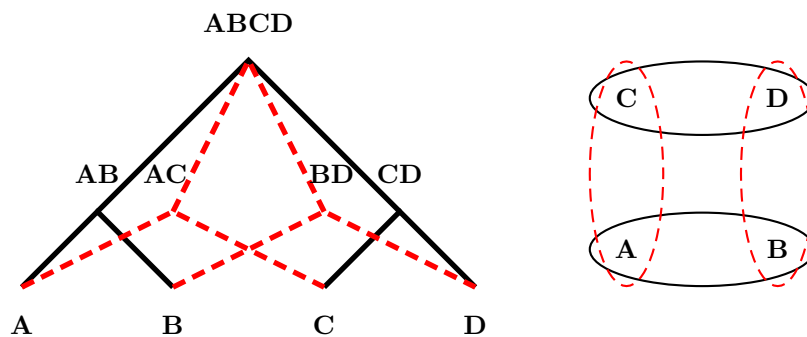


Fig. 6: Cluster network  $\mathcal{C}_2$

### 3 Affine Bruhat–Tits buildings and cluster networks

In the present section we will show that for the network of balls in  $\mathbb{Q}_p^d$  there exists a natural structure of simplicial complex which is related to the affine Bruhat–Tits building. Discussion of buildings one can find in particular in [6]. For discussion of  $p$ -adic geometry (in particular lattices) see [7].

**Affine Bruhat–Tits building.** Vertices of the building are equivalence classes of lattices. A lattice in  $\mathbb{Q}_p^d$  is an open compact  $\mathbb{Z}_p$ -module in  $\mathbb{Q}_p^d$ . Any lattice can be put in the form

$$\oplus_{i=1}^d \mathbb{Z}_p e_i,$$

where  $\{e_i\}$  is a basis in  $\mathbb{Q}_p^d$ .

Two lattices are equivalent if one is a scalar multiple of the other.

Two lattices  $L_1$  and  $L_2$  are adjacent (connected by an edge) if some representatives from equivalence classes  $L_1$  and  $L_2$  satisfy

$$pL_1 \subset L_2 \subset L_1.$$

$k$  – 1-Simplices are defined as equivalence classes of  $k$  adjacent lattices, i.e. the chains

$$pL_k \subset L_1 \subset L_2 \subset \dots \subset L_k.$$

Here  $1 \leq k \leq d$ .

An apartment in the affine building is the subcomplex corresponding to a fixed basis  $\{e_i\}$  in  $\mathbb{Q}_p^d$  which contains the equivalence classes of lattices  $\oplus_{i=1}^d \mathbb{Z}_p p^{a_i} e_i$ ,  $a_i \in \mathbb{Z}$ .

**Multidimensional  $p$ -adic metric.** Let us consider a metric  $s_{q_1, \dots, q_d}(x, y)$  in  $\mathbb{Q}_p^d$  defined by the norm  $N_{q_1, \dots, q_d}(z)$

$$s_{q_1, \dots, q_d}(x, y) = N_{q_1, \dots, q_d}(x - y), \quad (1)$$

$$N_{q_1, \dots, q_d}(z) = \max_{i=1, \dots, d} (q_i |z_i|_p), \quad q_i \neq 0. \quad (2)$$

Dilations  $p^k \mathbb{Z}_p^d$ ,  $k \in \mathbb{Z}$  are balls with respect to all such norms  $N_{q_1, \dots, q_d}$  if  $p^{-1} < q_i \leq 1$ .

Here we use the following definition of norm for  $\mathbb{Q}_p^d$ : a norm is a function  $N(\cdot)$  on  $\mathbb{Q}_p^d$  taking values in  $[0, \infty)$  and satisfying the conditions:

- i) Nondegeneracy:  $N(x) = 0 \Leftrightarrow x = 0$ ;
- ii) Linearity:  $N(ax) = |a|_p N(x)$ ,  $x \in \mathbb{Q}_p^d$ ,  $a \in \mathbb{Q}_p$ ;
- iii) Strong triangle inequality:  $N(x + y) \leq \max[N(x), N(y)]$ .

A general norm ( $A$ -rotation of  $N_{q_1, \dots, q_d}$ ) is defined as

$$N_{q_1, \dots, q_d}^A(z) = N_{q_1, \dots, q_d}(Az), \quad (3)$$

where  $A$  is a matrix from  $\text{Gl}_d(\mathbb{Q}_p)$ . A metric  $s_{q_1, \dots, q_d}^A$  is defined by the norm  $N_{q_1, \dots, q_d}^A$  as above (1).

In particular for a norm  $N_{q_1, \dots, q_d}$  of the form (2) with

$$p^{-1} < q_1 < \dots < q_d \leq 1 \quad (4)$$

the set of intermediary  $N_{q_1, \dots, q_d}$ -balls between  $p\mathbb{Z}_p^d$  and  $\mathbb{Z}_p^d$  contains the balls

$$B_j = \mathbb{Z}_p \times \dots \times \mathbb{Z}_p \times p\mathbb{Z}_p \times \dots \times p\mathbb{Z}_p$$

with  $j$  components  $\mathbb{Z}_p$  and  $d - j$  components  $p\mathbb{Z}_p$ ,  $j = 0, \dots, d$ .

**Simplicial complex of balls.** Let us define a structure of simplicial complex on the network  $\mathcal{C}$  of balls with respect to the defined above family of metrics  $s_{q_1, \dots, q_d}^A$ ,  $A \in \text{Gl}_d(\mathbb{Q}_p)$ .

Let  $s$  be a metric from the described family and  $I$  a  $s$ -ball containing zero (a  $s$ -ball is a ball with respect to  $s$ , zero is a vector in  $\mathbb{Q}_p^d$  with zero coordinates). Then the dilation  $pI$  is also a  $s$ -ball. The (containing zero)  $s$ -balls  $I$  and  $J$  are adjacent if  $pI \subset J \subset I$ .  $k - 1$ -Simplices are defined as families of  $k$  adjacent  $s$ -balls

$$pI_k \subset I_1 \subset I_2 \subset \dots \subset I_k.$$

Let us consider the maximal sequence of nested intermediary  $s$ -balls between  $pI$  and  $I$ . If the parameters  $q_i$  of the norm are generic (any two parameters can not be made equal by multiplication by degrees of  $p$ , for example when the parameters satisfy (4)) then the above sequence contains  $d + 1$  balls and defines a  $d - 1$ -simplex.

General simplices in the simplicial complex of balls with respect to a metric  $s_{q_1, \dots, q_d}^A$  are defined as translations of simplices described above (translations as families of sets in  $\mathbb{Q}_p^d$ ).

The simplicial complex  $\mathcal{C}$  of balls with respect to the family  $\{s_{q_1, \dots, q_d}^A\}$  of metrics is defined as a union of complexes of balls for different metrics  $s_{q_1, \dots, q_d}^A$ . Here we identify  $s$ -ball and  $s'$ -ball which coincide as sets (and identify  $s$  and  $s'$ -simplices which coincide as sets of balls).

**Relation between norms in  $\mathbb{Q}_p^d$  and simplices in the affine building.**

Any ball with respect to a norm in  $\mathbb{Q}_p^d$  which contains zero is a lattice. This follows from the strong triangle inequality.

Let us consider in the defined above simplicial complex  $\mathcal{C}$  of balls the sub-complex  $\mathcal{C}_0$  of balls which contain zero. For any ball  $I$  containing zero a dilation  $p^k I$ ,  $k \in \mathbb{Z}$  is also a ball (with respect to the same norm). The same holds for simplices. Therefore the factor  $\mathcal{C}_0/\Gamma$  by the group of dilations by  $p^k$ ,  $k \in \mathbb{Z}$  is a simplicial complex.

There exists a natural simplicial map from the simplicial complex  $\mathcal{C}_0/\Gamma$  to the affine Bruhat-Tits building which put in correspondence to a ball the corresponding lattice.

The defined map is an embedding of the complex  $\mathcal{C}_0/\Gamma$  into the affine building. Let us show that this map is surjective (i.e. is an isomorphism of simplicial complexes).

We say that two norms are equivalent if they generate the same family of balls. Let us show that to any maximal simplex in the affine Bruhat–Tits building one can put in correspondence an equivalence class of norms in  $\mathbb{Q}_p^d$ .

Let  $L$  be a lattice in  $\mathbb{Q}_p^d$  and

$$pL = L_0 \subset L_1 \subset \dots \subset L_d = L \quad (5)$$

be a maximal sequence of (different) embedded lattices (equivalently, a maximal simplex in the affine building).

For any pair  $L_j \supset L_{j-1}$ ,  $j = 1, \dots, d$  of consecutive lattices in the above sequence let us choose an element  $f_j \in L_j$ ,  $f_j \notin L_{j-1}$ . This gives a set  $\{f_1, \dots, f_d\}$  of vectors in  $\mathbb{Q}_p^d$ . One has the following lemma.

**Lemma 1** 1) The defined above set  $\{f_1, \dots, f_d\}$  is a basis in  $\mathbb{Q}_p^d$ ;  
2) The lattices  $L_j$  have the form

$$L_j = \oplus_{i=1}^j \mathbb{Z}_p f_i \oplus \oplus_{i=j+1}^d p \mathbb{Z}_p f_i. \quad (6)$$

Let us introduce a norm in  $\mathbb{Q}_p^d$  as follows. Let us put in correspondence to lattices  $L_j$  from (5) some positive numbers  $q_j$ ,  $p^{-1} < q_1 < \dots < q_d \leq 1$ .

Let us define a function  $N(x)$  on  $\mathbb{Q}_p^d$  in the following way. For  $x \in L_j \setminus L_{j-1}$ ,  $j = 1, \dots, d$  we define  $N(x) = q_j$ . For  $x = 0$  we put  $N(x) = 0$ . We define  $N(\cdot)$  in all  $\mathbb{Q}_p^d$  using the condition  $N(p^k x) = p^{-k} N(x)$ ,  $k \in \mathbb{Z}$ .

**Lemma 2** The function  $N(\cdot)$  defined as above will be a norm in  $\mathbb{Q}_p^d$  satisfying the strong triangle inequality. The sequence (5) of lattices will be a maximal sequence of balls with respect to  $N(\cdot)$  which lie between the balls  $pL$  and  $L$ .

The introduced norm  $N$  belongs to the family (3) with the parameters satisfying (4). In particular, the matrix  $A$  can be chosen as the matrix which maps the basis  $\{f_1, \dots, f_d\}$  from the above lemma to the coordinate basis in  $\mathbb{Q}_p^d$  and  $N(\cdot) = N_{q_1, \dots, q_d}^A(\cdot)$ . Any two norms defined in this way will be equivalent (will generate the same set of balls).

We have constructed a norm of the form (3), (4) starting from a maximal simplex in the affine building. Analogously, let us consider for a norm  $N_{q_1, \dots, q_d}^A(\cdot)$  defined by (3), (4) the set of lattices (6) where the basis  $\{f_1, \dots, f_d\}$  is defined by the matrix  $A$  as above. This set defines a simplex in the affine building.

We have shown that there exists a one to one correspondence between the equivalence classes of norms of the form (3), (4) and maximal simplices in the affine Bruhat–Tits building.

In the above construction it is important to consider norms with generic parameters. Let us consider a general norm  $N$  of the form (2), (3) and take a  $N$ -ball  $L$ . It is possible (say if some parameters  $q_i$  in (2) are equal) that a set of intermediary balls between  $L$  and  $pL$  contains less than  $d+1$  balls and therefore can not define a  $d-1$ -simplex in the affine building.



## 4 General networks of clusters

In the present section we discuss a generalization to general cluster systems of the described in the previous sections cluster network  $\mathcal{C}$  related to a family of metrics. We introduce a structure of simplicial complex on this network and a definition of dimension for cluster systems.

Let  $X$  be a locally compact ultrametric space with some finite family of ultrametrics  $\mathbf{s}$  defined on  $X$ . Moreover, let, for any pair of metrics  $s, r \in \mathbf{s}$ , any  $s$ -ball be a finite union of  $r$ -balls.

The family  $\mathbf{s}$  of ultrametrics on  $X$  is compatible, if for any two balls,  $s$ -ball  $I$  and  $r$ -ball  $J$ ,  $s, r \in \mathbf{s}$ , the intersection  $I \cap J$  is a ball with respect to some ultrametric  $t \in \mathbf{s}$ .

We put in correspondence to a metric  $s \in \mathbf{s}$  the corresponding tree  $\mathcal{T}(X, s)$  of  $s$ -balls in  $X$ . Vertices of this tree are  $s$ -balls, two vertices are connected by edge if the corresponding balls are nested without intermediaries.

The graph  $\mathcal{C}(X, \mathbf{s})$  (the network of clusters in  $X$  with respect to the family  $\mathbf{s}$  of metrics) is a union of trees  $\mathcal{T}(X, s)$  of  $s$ -balls,  $s \in \mathbf{s}$ .

The set of vertices of  $\mathcal{C}(X, \mathbf{s})$  is the union of the sets of  $s$ -balls,  $s \in \mathbf{s}$ , edges connect  $s$ -balls (with the same  $s$ ) nested without intermediaries. The partial order in  $\mathcal{C}(X, \mathbf{s})$  is defined by inclusion of subsets in  $X$ . If some  $s$ -ball coincides with some  $r$ -ball as a set, they define the same vertex in  $\mathcal{C}(X, \mathbf{s})$ .

Let us define simplices in the network  $\mathcal{C}(X, \mathbf{s})$ . Let  $\mathbf{r} \subset \mathbf{s}$  be a subfamily of metrics in  $X$ . Let us fix some  $\mathbf{r}$ -ball  $I$  (i.e.  $I$  is a  $s$ -ball with respect to all  $s \in \mathbf{r}$ ). Let  $J$  be a smallest  $\mathbf{r}$ -ball which is strictly greater than  $I$ .

We define a simplex as a subset (of cardinality at least two) of the set of intermediary  $s$ -balls lying between  $I$  and  $J$  (for metrics  $s \in \mathbf{r}$ ).  $k - 1$ -Simplex will contain  $k$   $s$ -balls, in particular, an edge in  $\mathcal{T}(X, s)$  (a pair ball – maximal subball) will be a one-simplex.

We have defined a structure of simplicial complex on the tree  $\mathcal{T}(X, s)$ . The union over metrics  $s \in \mathbf{s}$  of these simplicial complexes defines a structure of simplicial complex on the network  $\mathcal{C}(X, \mathbf{s})$  (where we, as usual, identify vertices which correspond to balls coinciding as sets and identify simplices coinciding as sets of vertices).

The idea of this definition of simplices is taken from the  $p$ -adic case discussed in the previous section. In this case simplices in the affine Bruhat–Tits building are given by sets of balls lying between balls  $I$  and  $pI$ . If  $I$  is a ball with respect to several metrics  $s_{q_1, \dots, q_d}^A$  (say given by different rotations  $A$  or different order of indices  $q_j$ ) then  $pI$  will also be a ball with respect to the same metrics. Therefore the above definition of simplicial complex  $\mathcal{C}(X, \mathbf{s})$  generalizes the definition of the simplicial complex  $\mathcal{C}$  of balls in  $\mathbb{Q}_p^d$  of the previous section (instead of consideration of simplices as subsets of a sequence of nested balls between  $I$  and the dilation  $pI$  we consider cycles in the cluster network  $\mathcal{C}(X, \mathbf{s})$ , simplex is a subset of a path between the minimal and maximal vertices of a cycle).

**Dimension of clusters network.** Let us generalize the definition of dimen-

sion (the number of  $p$ -adic parameters) to general cluster networks.

Let  $\mathbf{r} \subset \mathbf{s}$  be a subfamily of metrics in  $X$ . Let  $I, J$  be a pair of  $\mathbf{r}$ -balls ( $s$ -balls for all  $s \in \mathbf{r}$ ), where  $I \subset J$  and  $J$  is a minimal  $\mathbf{r}$ -ball which contains  $I$ .

Let us consider the maximal  $s$ -simplex,  $s \in \mathbf{r}$  corresponding to  $I, J$  (i.e. the maximal sequence of nested balls lying between  $I$  and  $J$ ).

The rank of this simplex we call the  $\mathbf{r}$ -dimension for the pair  $I, J$ <sup>1</sup>.

In the  $p$ -adic case: the family  $\mathbf{r}$  is given by the set of metrics  $s_{q_1, \dots, q_d}^A$  with fixed matrix  $A$  and all possible reorderings of parameters  $p^{-1} < q_1 < \dots < q_d \leq 1$ . The above dimension is equal to the number  $d$  of  $p$ -adic parameters.

The introduced dimension is not equal to the VC (or Vapnik – Chervonenkis, or combinatorial) dimension [8].

**Applications to data analysis.** Clustering is a tool of data analysis with many applications, in particular to bioinformatics. The set  $X$  of data may be generated in a complex way, there may be some independent contributions in the data.

There should be some way to describe independencies in data at the level of networks of clusters. Classification trees (trees of clusters) describe the diversity of data, the multidimensional generalization should describe the situation where we have independent sources of diversity.

Dimension of a network of clusters will describe the number of sources of diversity.

**Applications to taxonomy for reticulate evolution.** Let us discuss the application of the described above cluster networks in bioinformatics. One of the problems in bioinformatics is a construction of phylogenetic classification trees by comparison of genetic markers (some subsequences in genomes). These trees are constructed with the help of clustering procedure using some metric for genetic markers.

We are interested in the case when we have several genetic markers. In this case the metric is not uniquely defined. This metric has the form of a sum of contributions from the different genetic markers

$$d(X, Y) = \sum_{j=1}^N w_j d_j(X_j, Y_j),$$

where  $w_j \geq 0$  are weights,  $X$  and  $Y$  are genomes,  $X_j$  and  $Y_j$  are genetic markers,  $d_j$  is the distance for the  $j$ -th genetic marker.

Different sets of weights generate the different classification trees. In particular, when only one weight  $d_j$  is non zero, this weight generates the classification tree for the corresponding genetic marker. Union of these trees sometimes is called the forest of life [9].

The observation is that for some genetic markers the corresponding classification trees are different because different genetic markers may have different evolution histories. The evolution is *reticulate* — some parts of a genome may

---

<sup>1</sup>In principle this rank may differ for different  $s \in \mathbf{r}$ , in this case we take the maximal rank.

have the different origin due to hybridization or horizontal gene transfer. In this case instead of phylogenetic trees one can consider *phylogenetic networks*. Mathematical methods of analysis of phylogenetic networks one can find in [10, 11].

Discussed in the present paper classification networks might give a general framework for construction of phylogenetic networks. In this approach instead of reproduction of the detailed genetic history of populations (which in general is not possible) one could use classification networks for coarse grained description of the evolution of ensembles of genes.

**Acknowledgments** This paper was partially supported by the Program "Modern problems of theoretical mathematics" of the Department of Mathematics of the Russian Academy of Sciences.

## References

- [1] S.Albeverio, S.V.Kozyrev, Clustering by hypergraphs and dimensionality of cluster systems, *p-Adic Numbers, Ultrametric Analysis and Applications*. 2012. V.4. No.3. P.167–178. arXiv:1204.5952v1
- [2] S.V. Kozyrev, Multidimensional clustering and hypergraphs, *Theoretical and Mathematical Physics*, 2010, V.164. no.3. P.1163–1168.
- [3] J. Benois-Pineau, A. Khrennikov, Significance Delta Reasoning with p-Adic Neural Networks: Application to Shot Change Detection in Video. *The Computer Journal*, **53** (4), 417-431 (2010).
- [4] J. Benois-Pineau, A. Khrennikov, N. V. Kotovich, Segmentation of images in *p*-adic and Euclidean metrics. *Doklady Mathematics*, **64**, N. 3, 450-455 (2001).
- [5] F. Murtagh, *Multidimensional Clustering Algorithms*, Physica–Verlag, Wurzburg, 1985.
- [6] P.B.Garrett, *Buildings and Classical Groups*, Chapman and Hall, 1997.
- [7] A.Weil, *Basic number theory*, Springer Verlag, New York, 1967.
- [8] V.N.Vapnik, *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
- [9] E.V.Koonin, *The Logic of Chance: The Nature and Origin of Biological Evolution*. FT Press, 2011.
- [10] D.H. Huson, R. Rupp, C. Scornavacca, *Phylogenetic Networks*, Cambridge University Press, 2010.
- [11] A. Dress, K.T. Huber, J. Koolen, V. Moulton, A. Spillner, *Basic Phylogenetic Combinatorics*, Cambridge University Press, 2012.